PostgreSQL High Availability (HA)

What it is and Solutions

Author: RadixTrie Open-Source Team

Defining High Availability (HA)

High availability is of great significance for mission-critical systems, where a service disruption may lead to adverse business impact, resulting in additional expenses or financial losses. Although high availability does not eliminate the threat of service disruption, it ensures that the IT team has taken all the necessary steps to ensure business continuity.

There are key attributes that define a high availability system. It is important that we define these attributes before assessing these solutions.

- There should not be **any** single point of failure in the system.
- Continuous health monitoring of backend servers/systems/databases/applications.
- Reliable failover procedures and processes in case of failure.

HA Benefits

1. Extend and Enhance Protection Across Your Infrastructure
   a. Provide uniform, automated protection for all applications without modifications to the application or guest operating system.
   b. Establish a consistent first line of defence for your entire IT infrastructure.
   c. Protect applications that have no other failover options that might otherwise be left unprotected.

2. Scalability
   a. Master-slave node relationship—Replaces primary and secondary nodes. This new relationship model between nodes in a cluster enables availability actions to be coordinated by a single master node.
   b. Support for IPv6 networking—Enables an IT department in need of a larger "address" space to fully leverage its network infrastructure.
   c. Simple deployment mechanism—Promotes fast and easy completion of routine tasks such as deploying a High Availability agent and configuring High Availability functionality.

3. Reliability
   a. Elimination of external component dependencies—High Availability does not depend on DNS resolution. This reduces the likelihood that an external component outage will disrupt High Availability operations.
   b. Multiple communication paths—High Availability nodes within a cluster can communicate through the storage subsystem as well as over the management network. Multiple communication paths increase redundancy and enable better assessment of the health of a host and its virtual machines.

Comparing HA to Similar Systems

High Availability VS Fault Tolerance
While high-availability environments aim for 99.99% or above of system uptime, fault tolerance is focused on achieving absolute zero downtime.

High Availability VS Redundancy
The primary goal of high availability is to ensure system uptime even in the event of a failure. Redundancy, on the other hand, is the use of additional software or hardware to be used as a backup in the event that the main software or hardware fails.

High Availability VS Disaster Recovery
High availability is a concept wherein we eliminate single points of failure to ensure minimal service interruption.
On the other hand, disaster recovery is the process of getting a disrupted system back to an operational state after a service outage

HA for PostgreSQL in a Cluster

We need to redefine this in a database context. Using these criteria to evaluate single primary high availability solutions for PostgreSQL.

- Continuous server health monitoring.
- A notification mechanism that can inform administrators about the server or cluster degradation
- A way of connecting to the primary database server at all times.
- In case of primary PostgreSQL server failure, the most suitable standby server is promoted seamlessly.
- In case of standby failure, the system can restart the service or notify an administrator.
- System has fencing ability:
- In case of network isolation, the system is able to avoid split-brain scenarios.
- In case of replication does not happen on a given standby, the system is able to identify that, notify the administrator, and remove the faulty system from the cluster.

A Cluster is a collection of hosts that work together and are viewed as a single entity.

Single Cluster:
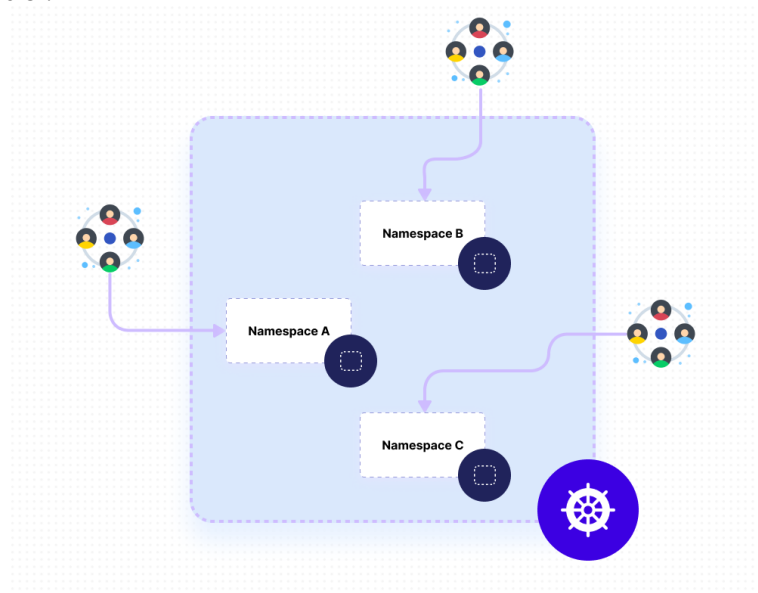
Multiple Clusters:
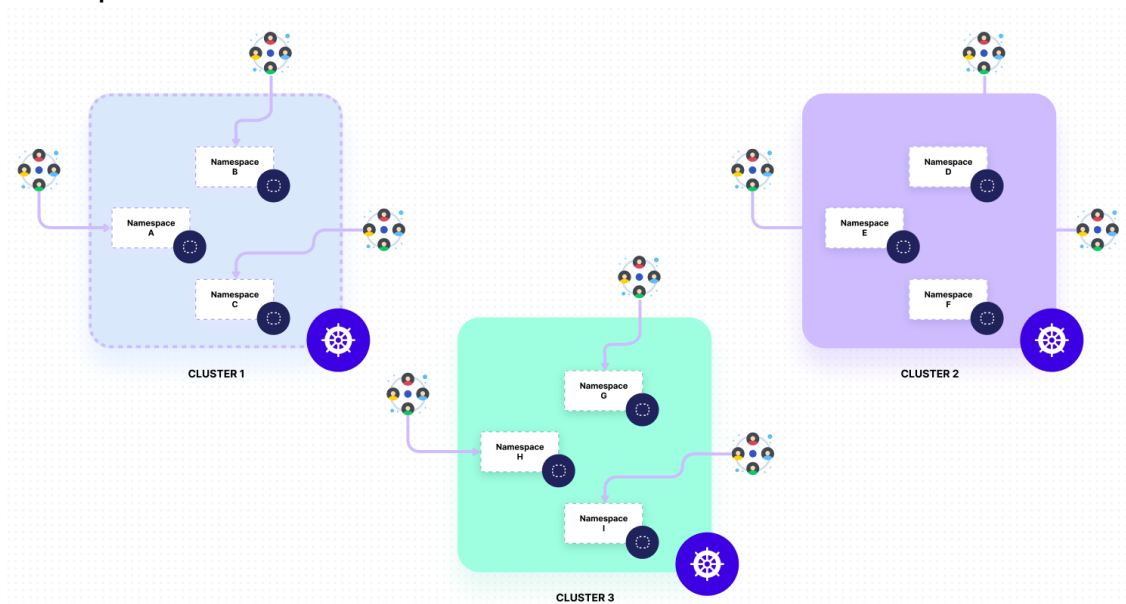
Types of HA
1. Infrastructure HA
2. DR / Failover
3. Application HA
4. Multi-geo-location application HA

HA Solutions

Tool Name	Tool Type
PgPool-II	Connection Pooler
PostgreSQL Automatic Failover (PAF)	Open Cluster Framework (OCF)
Replication Manager (RepMgr)	Replication
Patroni	All Purpose Solution
PgBouncer	

Best Practices
1. Assess your business requirements
2. Set Recovery Point Objective (RPO) and Recovery Time Objective (RTO) based on your expected availability percentage
3. Prepare a thorough disaster recovery program
4. Introduce redundancy strategically
5. Understand the metrics

Summary

In a nutshell, high availability implies there is no single point of failure. Everything from load balancer, firewall and router, to reverse proxy and monitory systems, is completely redundant at both network as well as application level, guaranteeing the highest level of service availability.

Although not a core requirement, but good for disaster management is to distribute resources geographically to avoid region-specific failures. Customer satisfaction often relies on whether or not customers can access a product or service when they need to and whether or not they can depend on it to work. High availability architecture ensures that a website, application, or server continues to function through different demand loads and failure types.